New Clustering Procedures in Respect of Practical Realization

Koch György and József Dombi

Clustering is one of the most frequently used statistical analyzing method. Its aim is to divide a set $A = \{x_1, x_2..., x_N\}$ of N objects into C categories called clusters in such a way that objects belonging to the same cluster should be as similar to each other as possible while objects in different clusters should be dissimilar.

Since there is no widely acceptable mathematical formula for similarity or dissimilarity, clustering problem has no universally usable optimal solution. The goodness of a clustering algorithm can only be tested empirically.

Depending on the given information there are two categories of clustering methods.

• If the parameters of the objects are given, the so-called coordinate based clustering algorithms can be used. These methods handle the objects like points in the m-dimensional space. Fastest algorithms in this group have $O(n)$ complexity which makes them suitable to process large databases. Well-known methods from this class are Kohonen Clustering Networks(KCN) and Fuzzy C- Means(FCM). We developed a new coordinate based clustering algorithm which has $O(n)$ complexity too and has several advantages over the previous methods. This algorithm is called Shepherd Method(SM) because of its behaviour. In Shepherd Method we iteratively separate the clusters with hyper planes (like a special Voronoi diagram) so it can be used easily for classification too. We also worked out a method for SM to automatically determine the number of clusters.

• Distance based clustering algorithms are the second group of clustering algorithms. In this case we do not know the exact parameters of the objects only their distances from each other or the similarity/dissimilarity between them. Hence this similarity is given by a matrix of size n*n, these algorithms have $O(n^2)$ time complexity, but these methods usually give better results than algorithms from the coordinate based group. One of the best distance based method is Chameleon which was published in 1999. We developed a new distance based method, which gives the same or even better results that Chameleon does. It is worth to mention that we achieved to handle relatively small amount of objects. Our method is called SmallSteps because it tries to find connected graphs which have egdes with a maximum weight which is computed on the environments of the objects. SmallSteps is capable to detect clusters with different shapes, sizes or densities and it is able to automatically determine the number of clusters needed and has the speciality to divide clusters into sub-clusters.